FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Solução de armazenamento na nuvem para dados científicos na U. Porto

José Pedro Marques Barbosa

VERSÃO DE TRABALHO

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Maria Cristina de Carvalho Alves Ribeiro

Co-orientador: João António Correia Lopes

22 de Julho de 2012

Solução de armazenamento na nuvem para dados científicos na U. Porto

José Pedro Marques Barbosa

Mestrado Integrado em Engenharia Informática e Computação

22 de Julho de 2012

Resumo

As novas tecnologias digitais impulsionaram a geração de dados científicos, por um lado devido à capacidade de armazenamento digital, por outro devido à evolução de métodos e ferramentas de investigação.

A partilha de dados científicos é essencial no processo de novas descobertas e é a forma pela qual os investigadores ganham reputação pelo seu trabalho.

Várias entidades, como universidades e comunidades de investigação, disponibilizam serviços e infraestruturas para auxiliar a curadoria e partilha de dados. Contudo, o processo de depósito de dados nestas infraestruturas é complexo e exige esforço conjunto de curadores e investigadores, pelo que muitos dos dados de investigação gerados acabam por não chegar a estes repositórios.

Este relatório descreve o estado da arte nestes campos e, também, uma proposta de dissertação cujo objetivo é aproximar os investigadores do processo de curadoria através de uma plataforma baseada em serviços na nuvem, agilizando o processo de submissão de dados nos repositórios.

A este serviço pretende-se aplicar alguns conceitos (como a simplicidade) de aplicações de armazenamento na nuvem familiares aos investigadores, como por exemplo a Dropbox, de forma a simplificar todo o processo de depósito de dados científicos em repositórios.

A solução deve ser um sistema independente que permita a integração com qualquer repositório de dados. Deve, também, permitir a anotação de dados por parte dos investigadores recorrendo, para isso, a um sistema de anotação.

Existe um conjunto de investigadores disponíveis a colaborar com este projeto que irão testar e validar a solução. Futuramente, com a implementação desta plataforma, é esperado um aumento da submissão de dados em repositórios de dados científicos, especialmente em ambientes multidisciplinares.

Abstract

New digital technologies have led to an increased scientific data growth, due to digital storage capacity and the development of new methods and research tools.

The sharing of scientific data is essential in the discovery process and represents the dominant means by which researchers can earn credits for their work.

Several entities, such as universities and research communities, are providing some infrastructures and services to improve data sharing and curation. However, the proccess of uploading data into these infrastructures is complex and demands joint efforts from curators and researchers, whereby much of the generated research data don't get to be uploaded to these repositories.

This report describes the state of the art in these fields and a thesis proposal whose goal is to bring researchers into the data curation process through a cloud storage service, improving the process of data submission in scientific repositories.

Some cloud storage applications concepts (such as simplicity as seen in Dropbox) should be applied to this service, in order to simplify the overall process of deposit data into scientific repositories.

The solution should be an independent system that enables integration with any data repository. It should also allow data annotation by researchers, making use of an annotation system.

There is a group of researchers collaborating with this project that will test and validate the solution. It is expected, in a near future, an increase of scientific data submissions into repositiries, specially in multidisciplinary environments.

Conteúdo

CONTEÚDO

Lista de Figuras

LISTA DE FIGURAS

Abreviaturas e Símbolos

UP	Universidade do Porto
HTTP	Hypertext Transfer Protocol
FTP	File Transfer Protocol
WebDav	Web Distributed Authoring and Versioning
API	Application Programming Interface

Capítulo 1

Introdução

1.1 Contexto

Nas últimas décadas tem-se assistido a um aumento significativo na quantidade e complexidade de dados de investigação, muito devido à evolução de métodos, instrumentos e ferramentas aliados à capacidade de armazenamento digital atual [Fer06].

A preservação e partilha de dados científicos é essencial para reutilização por parte da comunidade. Contudo, ainda existem muitos investigadores que não partilham os seus dados de investigação por vários motivos, entre os quais a complexidade do processo de depósito de dados em repositórios [RR, CPRR11].

A Universidade do Porto desenvolveu um projeto denominado *UPData* cujo objetivo foi o estudo das necessidades de curadoria multidisciplinar de dados científicos. No âmbito deste projeto foi desenvolvido um prototipo de um repositório de dados que se encontra atualmente a ser utilizado por investigadores.

O projeto proposto neste relatório pretende aproximar os investigadores do processo de curadoria, facilitando o processo de depósito de dados científicos em repositórios através de um serviço baseado em armazenamento na nuvem.

1.2 Motivação e Objectivos

O trabalho de dissertação proposto consiste no estudo do armazenamento de dados de investigação na nuvem. Será implementada uma solução de demonstração do conceito para de investigadores da Universidade do Porto (UP).

O depósito de dados no atual repositório da UP é realizado através de contacto direto entre o curador e o investigador. Assim, o objetivo principal desta dissertação é aproximar os investigadores do processo de curadoria através desta plataforma baseada em serviços na nuvem, agilizando o processo de submissão de dados no repositório de dados.

Questões de usabilidade, segurança e privacidade serão estudadas e abordadas, com o intuito de definir uma experiência de navegação o mais simples e segura possível. A plataforma a desenvolver deverá integrar com um sistema de anotação a ser desenvolvido atualmente para que todos os dados de investigação possam ser devidamente anotados pelos investigadores.

Este projeto visa melhorar o atual fluxo de inserção de dados no repositório da UP, com o objetivo de incentivar investigadores à submissão dos seus dados de investigação. Consequentemente, espera-se uma maior visibilidade da UP no que diz respeito à partilha de dados na *web*.

Existe, atualmente, uma equipa de investigadores a colaborar com este projeto que, futuramente, irão validar a solução proposta.

1.3 Estrutura da Dissertação

Além do capítulo de introdução, este relatório inclui quatro outros capítulos. No Capítulo 2 é apresentado o estado da arte que se encontra dividido em duas partes. A primeira aborda o tema da curadoria de dados e são apresentados alguns repositórios de dados relevantes, a segunda abrange serviços e tecnologias na área de armazenamento na nuvem. O Capítulo 3 descreve o problema no domínio deste projeto, bem como a sua contribuição para o estado da arte. O Capítulo 4 contém uma visão de alto nível sobre a proposta de solução ao problema. Inicialmente, são apresentados os requisitos da solução, seguidos da proposta de solução que engloba as decisões tecnológicas e de *design* tomadas, bem como a metodologia e prova do conceito. Finalmente, no Capítulo 5 é apresentado o plano de trabalho relativo aos próximos meses.

Capítulo 2

Estado da Arte

O presente capítulo é dedicado ao estado da arte e está subdividido em duas partes. Primeiramente, será dada uma visão sobre a investigação, as suas necessidades, a análise de repositórios e sua abordagem à curadoria. Por sua vez, a segunda parte incide na análise de alguns produtos idênticos à solução proposta nesta dissertação e uma análise a algumas tecnologias utilizáveis no âmbito deste projeto.

2.1 Dados Científicos e Repositórios de Dados

2.1.1 Dados Científicos: Acesso e Preservação

Segundo Galileu, a ciência moderna baseia-se na observação e experimentação. No decorrer da investigação torna-se vital que os investigadores registem informações pertinentes acerca das observações, ensaios e experiências realizados. Mais recentemente tem-se assistido a uma crescente preocupação na preservação desses registos, ou dados científicos, que permitem a contextualização e compreensão dos mesmos, por parte de outros investigadores [Fer06].

Nas duas últimas décadas, tem-se assistido a um aumento significativo na quantidade e complexidade destes dados devido à constante evolução de métodos, instrumentos e ferramentas aliados à capacidade de armazenamento digital atual. Assim, a ciência assume uma nova dimensão: ciência intensamente baseada em dados[1] [RSR+10].

No contexto de investigação, *datasets* são coleções de dados, geralmente representados sob forma tabular [RR]. Cada coluna representa uma variável e cada linha representa um membro do *dataset*. Um exemplo de um *dataset* relativo à área da gravimetria poderá ser visualizado na Figura 2.1

Os *datasets*[2] devem ser estruturados e organizados de forma a registarem fatos relacionados num formato comum para serem facilmente acedidos e interpretados no futuro. A preservação

[1] Do Inglês: *data-intensive science*.
[2] Conjuntos de dados.

grav.gpstime	grav.latitude	grav.longitude	grav.height
488743.999839	38.76028045	-27.084165796	113.155
488744.99984	38.760280428	-27.084165802	113.158
488745.999842	38.760280441	-27.084165801	113.159
488746.999843	38.76028044	-27.084165808	113.158
488747.999844	38.760280415	-27.084165805	113.158
488748.999845	38.760280466	-27.084165815	113.153

Figura 2.1: Exemplo de *dataset* na área da gravimetria.

destes *datasets* é, assim, essencial para a reutilização por parte da comunidade. Segundo a *UK e-Science*, "a partilha de dados e de recursos serão a chave para a resolução dos novos problemas da ciência e da engenharia" [Hey03].

A consciencialização da necessidade de partilha de *datasets* é cada vez maior por parte das entidades científicas, universidades e autoridades governamentais. Um exemplo disso, são dados públicos sobre cidades (descrições, limites das fronteiras, pontos de interesse, entre outros) que, desde 2009, têm vindo a ser disponibilizados pelos Estados Unidos da América com o objetivo de poderem ser usados e acedidos facilmente por qualquer pessoa [SL12]. Como consequência desta iniciativa surgem várias aplicações, principalmente, móveis e *web* que utilizam estes dados em vários contextos.

Cada vez mais os investigadores estão envolvidos no processo de partilha dos seus dados científicos, devido às vantagens oferecidas a longo prazo. Este processo concede maior dimensão, visibilidade e eventualmente continuidade ao trabalho desenvolvido. Contudo, ainda existem muitos investigadores que não partilham os seus dados por vários motivos. Nesse sentido, as instituições de investigação têm vindo a criar políticas institucionais que induzem o depósito de dados em repositórios adequados e incentivam à partilha [RSR+10].

A partilha de *datasets* na *web* tem sido impulsionada através de repositórios de dados disponibilizados por várias entidades. Aliados a estes repositórios encontram-se temáticas associadas à curadoria e gestão de dados que serão abordadas na próxima secção.

2.1.2 Curadoria de Dados

A curadoria de dados envolve manutenção, preservação e enriquecimento de dados de investigação, durante o seu ciclo de vida [Cen12]. A sua gestão ativa reduz o risco de obsolescência e diminui possíveis incoerências. Outro objetivo da curadoria é reduzir a duplicação de dados, permitindo, a longo prazo, uma pesquisa de qualidade .

Em 2005 foram produzidos 150 *exabytes* de dados, e, como é possível verificar na Figura 2.2, desde 2007 que a quantidade de dados produzidos pela sociedade deixou de ser suportada pela capacidade de armazenamento disponível [Eco10a]. Assim sendo, a curadoria surge, também, como um filtro ao crescimento exponencial da geração de dados nos últimos anos [Eco10b].

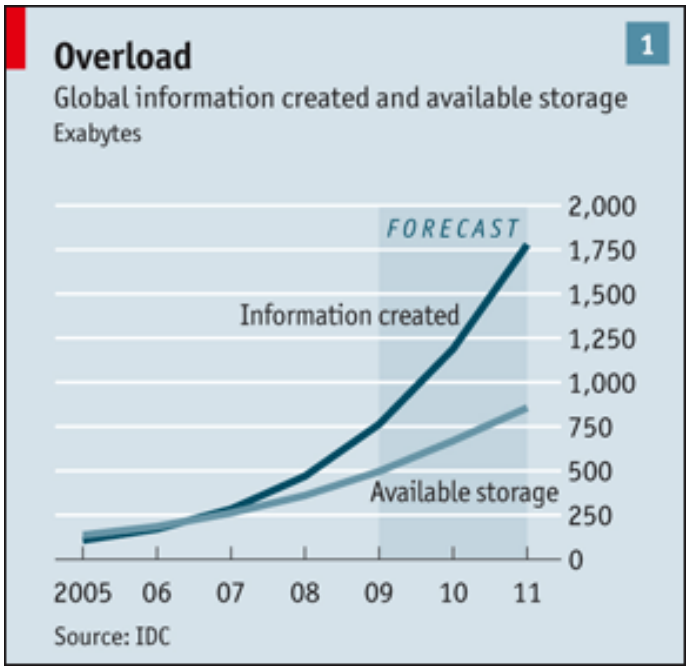

Figura 2.2: Dados criados e armazenamento disponível nos últimos anos [Eco10a].

Tipicamente existem dois tipos de pessoas envolvidas no processo de curadoria: investigadores e curadores. Um curador é responsável por colocar dados de investigação nos repositórios. Este lida com documentos de outras pessoas e debruça-se, essencialmente, na preservação, indexação e classificação para que possa ser acedido facilmente, através da Internet. O papel do investigador é efetuar publicações das suas investigações e experiências, dando importância à organização e anotação dos seus dados [BB04], na medida em que a preservação é assegurada por curadores. Tradicionalmente, estes atores interagem durante todo o processo de depósito de dados nos repositórios, de forma a minimizar alguns dos problemas associados ao processo.

De seguida, serão apresentados os problemas mais comuns:

- Tipicamente os *datasets* são imutáveis. E se estes mudarem frequentemente? [BB04]
- Os repositórios multidisciplinares incluem dados de áreas bastantes distintas. Nesse caso, será um curador capaz de curar todo o tipo de dados?
- Como selecionar os *datasets* para depositar? É pertinente preservar todos os dados gerados pelas comunidades? [RSR+10]

Cada vez mais, os responsáveis por repositórios de dados têm vindo a oferecer ferramentas, técnicas e serviços para auxiliar curadores e investigadores no processo de curadoria e resolver alguns dos problemas acima descritos [RSR+10]. Numa tentativa de minimizar problemas, em projetos de grande dimensão, é comum incluir um curador na equipa de investigação.

As vantagens da curadoria só se fazem sentir a longo prazo, como tal, nem sempre é fácil sensibilizar a comunidade a contribuir neste processo. A DCC [3],um centro de dados Britânico, implementou um conjunto de atividades com o objetivo de catalisar investigações inovadoras e partilhar resultados. As iniciativas incidiram, principalmente, em [BGAT]:

- Promover a necessidade de curadoria entre as comunidades de cientistas e investigadores;
- Prestar serviços, com o intuito de auxiliar e facilitar a curadoria;
- Realizar ações de partilha de experiências e conhecimento em curadoria entre diversas áreas de conhecimento;
- Desenvolver tecnologia de apoio à curadoria;
- Promover investigação na área de curadoria, inovando com novas técnicas e serviços a disponibilizar à comunidade.

O resultado deste tipo de iniciativas, a longo prazo, tem tido sucesso em várias universidades e comunidades de investigação [BGAT], contudo nem todas as comunidades têm recursos nem serviços para atuar e inovar nesta área [RSR+10].

Segundo um estudo sobre diferenças disciplinares na área da curadoria [Lyo10], torna-se necessário desenvolver estratégias para disciplinas dissemelhantes, pois a abordagem tradicional não será suficiente para responder às necessidades dos investigadores nas diversas áreas. É precisamente este um dos principais desafios de universidades detentoras de infraestruturas de curadoria de dados científicos.

2.1.3 Repositórios de Dados: Abordagens à Curadoria

A fraca robustez das atuais infraestruturas é uma preocupação para as organizações que lidam com grandes quantidades de dados científicos. Serão considerados quatro cenários distintos que representam a abordagem à curadoria de dados, por parte destas organizações. Esta subdivisão foi baseada no relatório "Os Repositórios De Dados Científicos: Estado Da Arte" [RSR+10].

Curadoria por Investigadores ou Técnicos

É comum este cenário em universidades ou centros de investigação em que não existe nenhuma política institucional para a curadoria. Isto acontece, mais precisamente, em grupos que necessitem de componentes de processamento de dados em áreas pouco exploradas a nível de formatos normalizados.

Os sistemas informáticos destas infraestruturas são, geralmente, sólidos e garantem que os dados se mantêm ativos, apesar dos dados não serem salvaguardados sistematicamente.

[3] *The Digital Curation Centre.*

Como exemplos de repositórios temos o CAVA[4], "Human Communication: an Audiovisual Archive" e o OASIS[5], "Open Access Series of Imaging Studies". Estes repositórios foram desenvolvidos e mantidos através de financiamentos de investigação [RSR+10].

Curadoria por Organizações Científicas

Este cenário envolve, geralmente, um trabalho em conjunto de universidades e instituições cientificas que fornecem serviços de acesso dentro de uma comunidade cientifica. Por norma estas associações cientificas desenvolvem ou contratam infraestruturas de curadoria de dados [RSR+10].

Os dados contidos nestes repositórios são de áreas específicas e é verificada uma descrição mais especializada. Normalmente estes dados são restritos à comunidade em que se inserem, no entanto algumas associações tornam os seus dados públicos.

Como exemplo deste cenário temos o repositório Holandês DANS, "Data Archiving and Networked Service"[6], na área das artes, humanidades e ciências sociais, suportado pela KNAW [7] e NWO[8]. O objetivo deste repositório é manter os dados acessíveis permanentemente dando especial foco ao histórico de dados [Rom10]. Naturalmente, este repositório disponibiliza um conjunto de serviços para depósito, análise e descarga de dados.

Outros exemplos neste cenário são os repositórios ICPSR, "Inter University Consortium for Political and Social Research", *UK Data Archive*, e NCBI, "National Center for Biotechnology Information". Este último, na área da biotecnologia e biomedicina, permite aos investigadores utilizarem diretamente um conjunto de bases de dados especializadas, pelo que pressupõe alguma familiaridade com o tema e suas representações.

Curadoria por Universidades ou Centros de Investigação

Este cenário em todo semelhante à abordagem referida anteriormente mas inclui dados de diversas áreas devido a serem iniciativas de universidades ou centros de investigação. O objetivo principal destes repositórios é dar visibilidade às instituições participantes. A curadoria é abordada de várias formas nestes cenários, sendo mais comum a utilização de bibliotecas e centros de computação. Um dos maiores problemas neste panorama é a inexistência de descrições especializadas para os conjuntos de dados das diversas áreas.

Devido à variedade de áreas com que estes repositórios lidam, é importante manter proximidade entre curadores e investigadores de forma a assegurar o sucesso na curadoria de dados nas bibliotecas. Por norma existem infraestruturas duráveis que permitem registar, descrever, pesquisar e aceder aos dados científicos.

[4] http://www.jisc.ac.uk/whatwedo/programmes/inf11/sue2/cava.aspx
[5] http://www.oasis-brains.org/
[6] http://www.dans.knaw.nl/
[7] Royal Netherlands Academy of Arts and Sciences.
[8] Netherlands Organization for Scientific Research

O repositório escocês *Datashare*[9], da Universidade de Edimburgo, é um exemplo deste cenário gerido por uma biblioteca. Isto tudo é possível em parceria com o EDINA[10], um centro de dados académicos que disponibiliza serviços à comunidade [Dat07].

Curadoria por Organismos Oficiais

Este cenário surge em países onde a sensibilidade à curadoria de dados científicos é maior, onde estas iniciativas partem de organismos de gestão da ciência do país que financiam estes projetos. A distância entre o serviço e os investigadores acaba por ser uma fragilidade desta abordagem que, aliada aos dados de diversas áreas incluídos nestes repositórios, acaba por prejudicar a descrição especializada dos dados [RSR+10].

Um exemplo deste cenário é o ANDS[11], "Australian National Data Service, financiado por diversos organismos oficiais Australianos como o objectivo de dar visibilidade dos seus dados cientificos na *web*. Outros exemplos igualmente importantes são o DataONE[12] e o NBII[13], este último que recentemente foi terminado devido a cortes de orçamento do governo dos Estados Unidos da América.

2.1.4 UPData: Curadoria na Universidade do Porto

A Universidade do Porto desenvolveu um projeto denominado *UPData* que tem como determinar as principais necessidades de curadoria de dados científicos em diversas áreas de investigação pertencentes à instituição. Atualmente, existe uma equipa de investigadores a cooperar com este projeto [RR, CPRR11].

Existe já um protótipo de um repositório de dados experimental a ser utilizado pela equipa de investigadores. Este protótipo é uma extensão à plataforma *open-source DSpace*[14]. Esta plataforma permite a diversas organizações instalar facilmente um repositório de dados, com facilidade de personalização [DSp12].

2.2 Armazenamento na Nuvem

A utilização do armazenamento na nuvem tem ganho popularidade desde a década de 90, com o aumento de largura de banda da Internet. Desde então, tem-se assistido cada vez mais a uma massificação deste conceito [Moh12].

O armazenamento na nuvem é um modelo de armazenamento *online* onde os dados são mantidos, geridos e salvaguardados remotamente e são disponibilizados aos utilizadores, através da Internet [Rou11].

[9] http://datashare.is.ed.ac.uk/
[10] http://edina.ac.uk/
[11] http://www.ands.org.au/
[12] https://www.dataone.org/
[13] https://www.dataone.org/
[14] http://www.dspace.org/

Os dados, a partir deste modelo, são armazenados em múltiplos servidores em vez de servidores dedicados, tipicamente empregues em redes tradicionais de armazenamento de dados. A localização dos ficheiros pode mudar a qualquer momento, visto que o sistema gere dinamicamente o espaço disponível nos vários servidores e balanceia o armazenamento, utilizando algoritmos de otimização. Contudo, apesar desta localização variável, o utilizador vê os ficheiros numa localização "estática", sendo-lhe permitida a gestão dos seus dados como se este estivesse a utilizar o seu próprio computador [WPG+10].

O acesso a estes serviços pode ser efetuado através de uma *interface web*, de API e, em alguns casos, é fornecido o acesso através de protocolos de comunicação como FTP (File Transfer Protocol) e WebDav (Web Distributed Authoring and Versioning). Este tópico será discutido mais detalhadamente na Secção 2.2.1.1.

Existem vários tipos de sistemas de armazenamento na nuvem, alguns com um foco específico, como por exemplo, arquivar emails, outros que lidam com todo o tipo de dados e permitem a sua gestão remota. A arquitetura típica destes sistemas inclui um servidor de controlo e vários servidores de armazenamento interligados, tal como demonstrado na Figura 2.3.

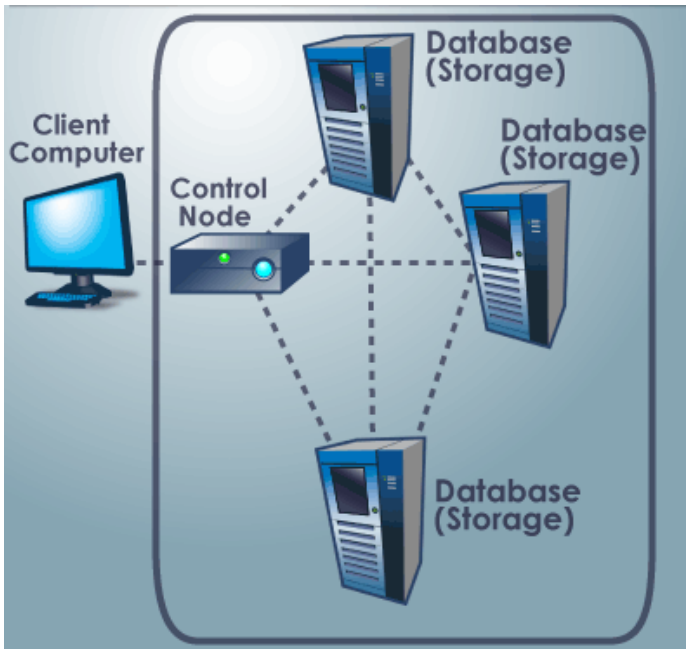

Figura 2.3: Típica arquitetura de sistemas de armazenamento na nuvem [WPG+10]

Tipicamente, o armazenamento na nuvem é mais barato do que a instalação e manutenção de servidores dedicados. Por este motivo, várias empresas têm adotado este modelo como resposta ao crescimento exponencial da geração de documentos. Para além de vantagens financeiras, existe replicação de dados nestes sistemas. Por motivos de segurança, os dados armazenamos na nuvem são replicados por vários servidores locais, assim, se por algum motivo um servidor falhar, é garantida a persistência de dados [SCL].

Armazenamento na nuvem como um Serviço

O armazenamento na nuvem como um serviço[15] é um modelo de negócio onde grandes empresas alugam espaço das suas infraestruturas a outras empresas ou indivíduos.

Desde 2007, quando a Google propôs o modelo de armazenamento na nuvem formalmente, que têm emergido vários fornecedores deste tipo de serviços. Estes fornecedores alugam espaço de armazenamento baseado no custo por *gigabyte* e por largura de banda e garantem a manutenção e gestão das suas infraestruturas. Os seus sistemas devem obedecer a um conjunto muito rigoroso de requisitos, incluindo replicação e consistência de dados, garantindo disponibilidade, desempenho, confidencialidade e segurança [HZZS11].

Exemplos de serviços neste ramo são: Amazon EC2[16], Google App Engine[17], Microsoft Azure[18] e IBM Blue Cloud[19].

Várias empresas contratam este tipo de serviços especializados a terceiros por ser mais viável economicamente, face à contratação de técnicos para implementar e manter infraestruturas de armazenamento. Isto permite à organização a possibilidade de se focarem na sua área de negócio [RJKG10].

2.2.1 Aplicações e Serviços de Armazenamento

Existem várias empresas que aproveitam o mercado do armazenamento na nuvem para criarem o seu modelo de negócio.

Atualmente, um serviço muito comum é a oferta de armazenamento *online* que permite aos utilizadores a sincronização e envio de ficheiros para a nuvem. É oferecido acesso a este sistema de armazenamento através de um navegador *web* ou de aplicações dedicadas que, quando instaladas num computador pessoal, sincronizam um diretório local com o respetivo diretório na nuvem.

Como exemplos que ilustram estes serviços temos a Dropbox[20], Microsoft Skydrive[21], Google Drive[22] e Ubuntu One[23]. Geralmente, as entidades responsáveis por estes serviços gerem grandes armazéns de dados em localizações geográficas distintas, no entanto a Dropbox recorre aos serviço da Amazon S3[24] como fornecedor de armazenamento na nuvem.

A simplicidade destes sistemas é a chave para o seu sucesso. Em fevereiro de 2012 a Dropbox chegou aos 50 milhões de utilizadores e segundo Drew Houston, o seu cofundador, o seu êxito pode ser resumido numa palavra: nuvem [Mur12]. Houve, ainda, uma preocupação e esforço em "criar um produto simples e elegante que satisfizesse os utilizadores".

[15] Do inglês: *Storage as a Service(SaaS)*.
[16] http://www.amazon.com/ec2/
[17] http://code.google.com/appengine/
[18] http://www.windowsazure.com/
[19] http://www.ibm.com/ibm/cloud/
[20] http://www.dropbox.com/
[21] http://skydrive.live.com/
[22] http://drive.google.com/start
[23] http://one.ubuntu.com/
[24] http://aws.amazon.com/s3/

Dentro do mesmo conceito existem várias aplicações *open-source* que podem ser instaladas numa infraestrutura pessoal e, tal como os serviços já referidos, permitem o acesso ao sistema de ficheiros através da Internet. A ownCloud[25] e FTPbox[26] são exemplos destas aplicações.

A ownCloud permite a utilização de armazenamento do servidor local ou a utilização dos serviços Amazon S3 e Google App Engine. Este sistema dedicado pode ser acedido através de qualquer dispositivo com ligação à Internet através de um navegador *web* ou de aplicações dedicadas *open source* fornecidas pela ownCloud.

Desta forma, qualquer indivíduo ou entidade pode criar um sistema de armazenamento próprio baseado nesta plataforma, precisando apenas de instalar a aplicação servidor numa infraestrutura dedicada.

2.2.1.1 Acesso ao Armazenamento na Nuvem

O acesso a ficheiros armazenados em serviços de armazenamento tradicional, ou hospedado, é efetuado diretamente com o servidor, através de protocolos de comunicação como o FTP e WebDav. O WebDav é uma extensão do protocolo HTTP que detém métodos que permitem a anotação de ficheiros, a gestão controlos de acesso, a gestão de versões, e ainda, possui métodos que implementam operações sobre ficheiros como copiar, eliminar, mover e o envio de ficheiros para o repositório [DN]. Devido a estes métodos o é muito usado em aplicações colaborativas e em aplicações que envolvam a gestão de ficheiros remota.

O armazenamento na nuvem é uma evolução do armazenamento hospedado que oferece o acesso através de API sofisticadas que permitem a abstração destes protocolos [WPG+10]. No entanto, é frequente ser suportado WebDav e FTP nestes serviços com o objetivo de permitir a integração com aplicações externas que implementem estes protocolos.

A Figura 2.4 mostra a evolução do armazenamento na nuvem com base no armazenamento tradicional ou hospedado.

2.2.2 Vantagens e Desafios

Como conclusão, existem vários motivos para a o crescimento da popularidade do armazenamento na nuvem e para a sua viabilidade no negócio. De seguida, é apresentada uma lista de cinco benefícios chave na utilização deste tipo de armazenamento em aplicações que lhe tirem proveito [WPG+10].

- Aplicações que tirem partido de armazenamento na nuvem são mais fáceis de configurar e gerir do que as tradicionais. Toda a complexidade do sistema de armazenamento é da responsabilidade do fornecedor do serviço.

- Geralmente, o armazenamento na nuvem é mais viável economicamente visto eliminar custos associados a sistemas dedicados. Atingir a qualidade (em termos de escalabilidade,

[25] http://owncloud.org/
[26] http://ftpbox.org/

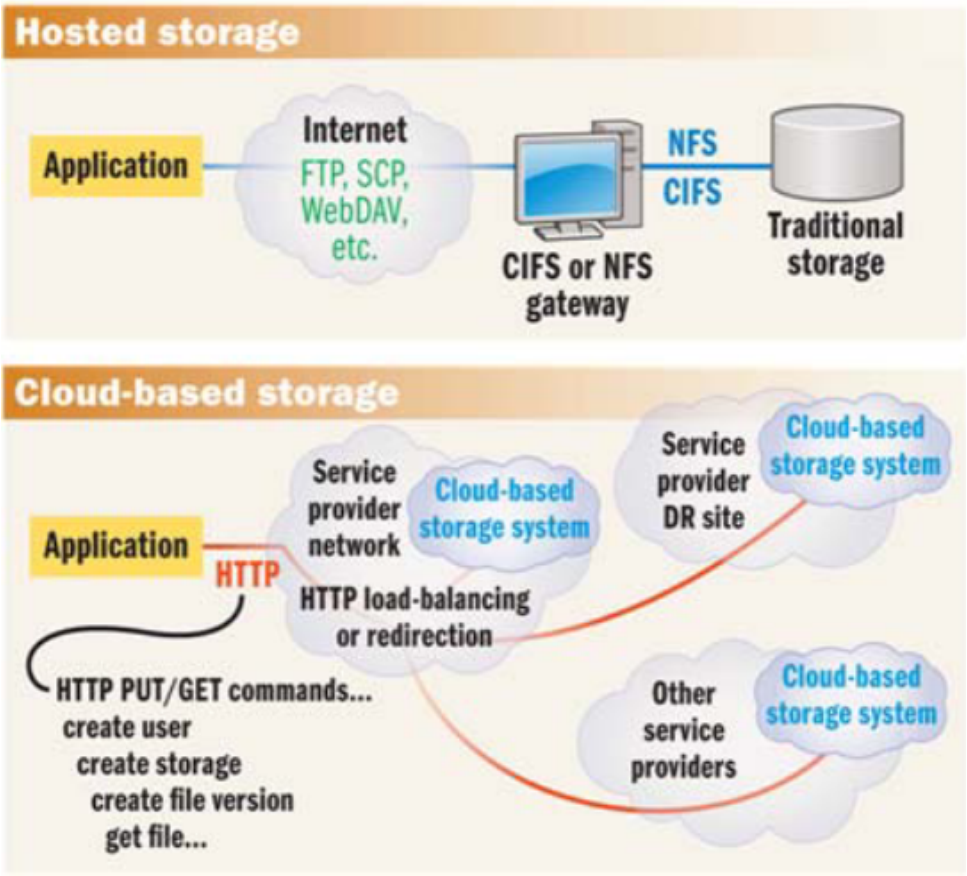

Figura 2.4: Evolução do acesso ao armazenamento na nuvem [WPG+10]

segurança, disponibilidade e persistência) dos serviços disponibilizados por detentores de grandes centros de dados é muito dispendiosa e praticamente inalcançável pela maior parte das empresas. Por este motivo, os serviços de armazenamento na nuvem geridos por terceiros são, na maior parte das vezes, compensatórios.

- As atualizações de *hardware* em sistemas tradicionais causam interrupções no acesso ao armazenamento. Com armazenamento na nuvem estas atualizações não serão visíveis ao utilizador final, eliminando as interrupções no serviço.

- Serviços de armazenamento na nuvem mantêm várias cópias de segurança em armazéns de dados situados em diversas zonas geográficas. Em caso de catástrofe natural numa zona, ou se simplesmente um armazém de dados falhar, não haverá perda de dados nem quebra no serviço.

- Planeamentos de armazenamento detalhados já não serão um problema. Serviços de armazenamento na nuvem são flexíveis e permitem armazenamento conforme as necessidades.

Recentemente, foi efetivado um estudo pela Faculdade de Economia e Ciência Políticas de Londres[27] e da Accenture[28] baseado em questionários a mais de mil executivos das tecnologias da informação (TI), bem como, em entrevistas a trinta e cinco prestadores de serviços na área do armazenamento na nuvem. Os entrevistados na área das TI mostraram-se mais cautelosos relativamente a prazos realistas para a implementação de nuvens do que os da área dos negócios, que estão mais interessados em soluções ágeis e rentáveis a curto prazo. Existem vários desafios na implementação de soluções de armazenamento na nuvem, esse é o motivo para a prudência dos executivos das TI [WVW12]. De seguida, são apresentados alguns desses desafios.

- Segurança: A segurança é um desafio comum a todas as aplicações acessíveis através da Internet, e não, um problema especifico do armazenamento na nuvem. Contudo, os grandes fornecedores de armazenamento na nuvem têm a capacidade de investir em *hardware* e *software* mais sofisticado para análise de comportamentos incomuns e deteção de vulnerabilidades, sendo que a resposta a ataques é, geralmente, bastante eficaz [Sab11].

- Aprisionamento tecnológico[29]: Atualmente, a mudança de fornecedor de serviços de armazenamento na nuvem implica custos substanciais [WVW12].

- Gestão da nuvem: Uma das grandes vantagens da utilização de serviços na nuvem disponibilizados por terceiros é a facilidade da sua atualização ou alteração, sem necessidade de intervenção interna. Esta funcionalidade disponibilizada por fornecedores destes serviços pode ser difícil de gerir [WVW12].

Os sistemas de armazenamento na nuvem são projetados para serem escaláveis e fáceis de manter. A contratação destes serviços a terceiros permite a abstração da sua complexidade, no entanto, deve ser bem planeada tendo em conta os diversos tópicos acima abordados.

[27] London School of Economics and Political Science
[28] http://www.accenture.com/
[29] Do inglês: Vendor lock-in.

Capítulo 3

Descrição do Problema

O trabalho de dissertação proposto consiste no estudo do armazenamento de dados de investigação na nuvem. Será implementada uma solução de demonstração do conceito para de investigadores da Universidade do Porto, disponibilizando uma API para acesso aos ficheiros armazenados no sistema.

Atualmente, o depósito de dados científicos no repositório experimental é realizado manualmente, através do contacto direto entre o curador e o investigador. O objetivo principal deste projeto é agilizar e automatizar este processo, e ainda aproximar os investigadores do processo de curadoria, oferecendo-lhes um serviço familiar de gestão e centralização de dados como incentivo à participação.

A elaboração de uma plataforma que resolva este problema e consiga responder às necessidades dos investigadores terá que ter em conta vários aspetos de *design*, com o intuito de manter a solução simples e facilmente usável, não comprometendo os seus objetivos.

A solução, *UPBox*, será um serviço que oferecerá aos investigadores armazenamento na nuvem[1], sincronização e gestão de ficheiros, estando estes acessíveis através de qualquer dispositivo com acesso à Internet. Este módulo da aplicação será semelhante a aplicações às quais os investigadores estão familiarizados, por exemplo a *Dropbox*[2], devido à sua simplicidade e usabilidade. Será desenvolvido um módulo para integração com um sistema de anotação, atualmente a ser desenvolvido no âmbito do projeto *UPData*, para permitir a anotação de dados por parte dos investigadores.

Como se pode verificar no fluxo de trabalho da Figura 3.1, a qualquer momento o investigador poderá submeter dados ou anotar dados submetidos, com auxílio do sistema de anotação. Mais tarde, quando o investigador definir um diretório pessoal como disponível para curadoria, o curador terá acesso a esse diretório e será responsável por colocá-lo num repositório apropriado.

[1] Ou armazenamento online.
[2] https://www.dropbox.com/

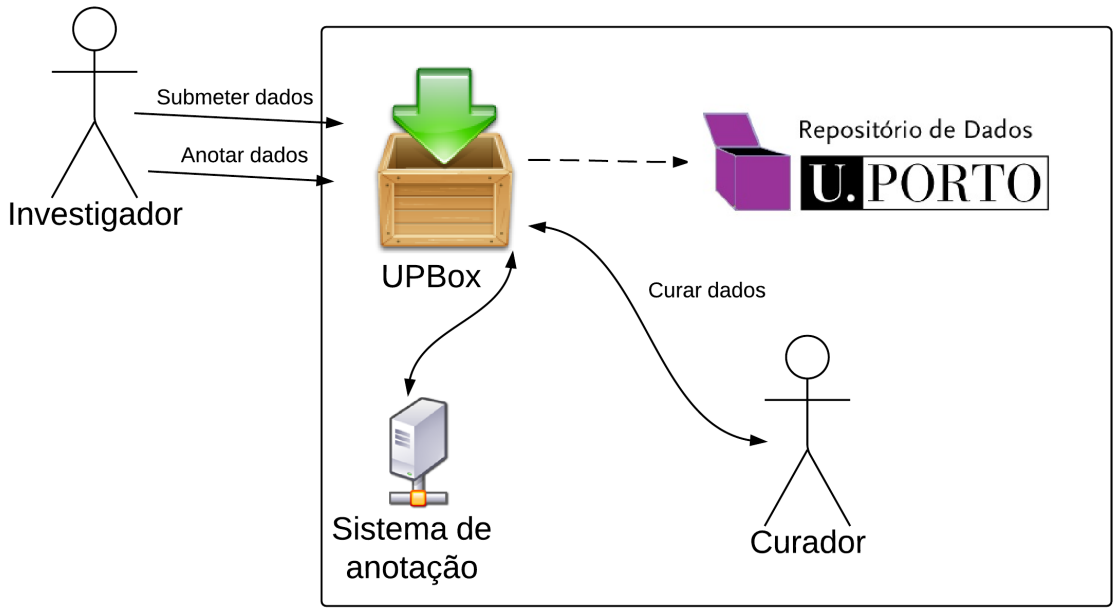

Figura 3.1: Fluxo de trabalho e integração do sistema.

Neste projeto, torna-se importante ter em conta questões de segurança e privacidade. Os dados dos investigadores no repositório serão privados e só poderão ser acedidos por curadores, mediante permissão de acesso por parte do investigador.

Existem várias abordagens ao sistema de ficheiros em soluções de armazenamento na nuvem, por isso vai ser necessário definir qual a mais adequada e quais os meios acesso aos ficheiros disponibilizados na UPBox.

Serão desenvolvidas três API: a primeira deverá permitir a integração com clientes externos para acesso e sincronização, a segunda permitirá o acesso a ficheiros, por parte de repositórios de preservação de dados. A terceira, e última, permitirá a comunicação com o sistema de anotação com o intuito de ser sugerida uma anotação para o conjunto de dados. No âmbito deste projeto, o repositório usado será o Repositório de Dados da UP [3].

[3] http://sciencedata.up.pt/dspace/

Capítulo 4

Perspectivas de Solução

O presente capítulo inclui a descrição da solução proposta. Serão apresentados os seus requisitos, assim como a proposta de solução, a análise da arquitetura e as decisões tecnológicas tomadas. Por fim, será referida a metodologia utilizada e será descrita a forma de provar o conceito.

4.1 Requisitos da Solução

Atendendo ao problema descrito no Capítulo 3 foram elaborados os requisitos funcionais e não funcionais da solução proposta. Como auxílio na definição dos requisitos foi tida em conta a auditoria feita na Universidade do Porto [REFP11], baseada em inquéritos a investigadores a cooperar com o *UPData*.

É importante oferecer um serviço vantajoso aos investigadores, do qual estes tirem o máximo proveito resolvam alguns dos seus problemas com grande simplicidade. Assim, a usabilidade e simplicidade são requisitos chave, não funcionais, da *UPBox*. Questões de segurança e privacidade têm que ser tidas em conta neste tipo de serviços, garantindo que dados pessoais de investigadores não estejam acessíveis a outros indivíduos.

Os requisitos funcionais podem ser consultados na Figura 4.1. Estes englobam funcionalidades de gestão de ficheiros remota, de autenticação e funcionalidades de auxílio à curadoria, por parte de curadores.

4.2 Proposta de Solução

Baseado no problema e nos requisitos elaborados, é possível visualizar a arquitetura da plataforma na Figura 4.2. De ressalvar que os componentes e conexões neste diagrama não pretendem ser exaustivos, podendo não traduzir o produto final, visto que os detalhes da arquitetura serão definidos numa fase posterior do projeto.

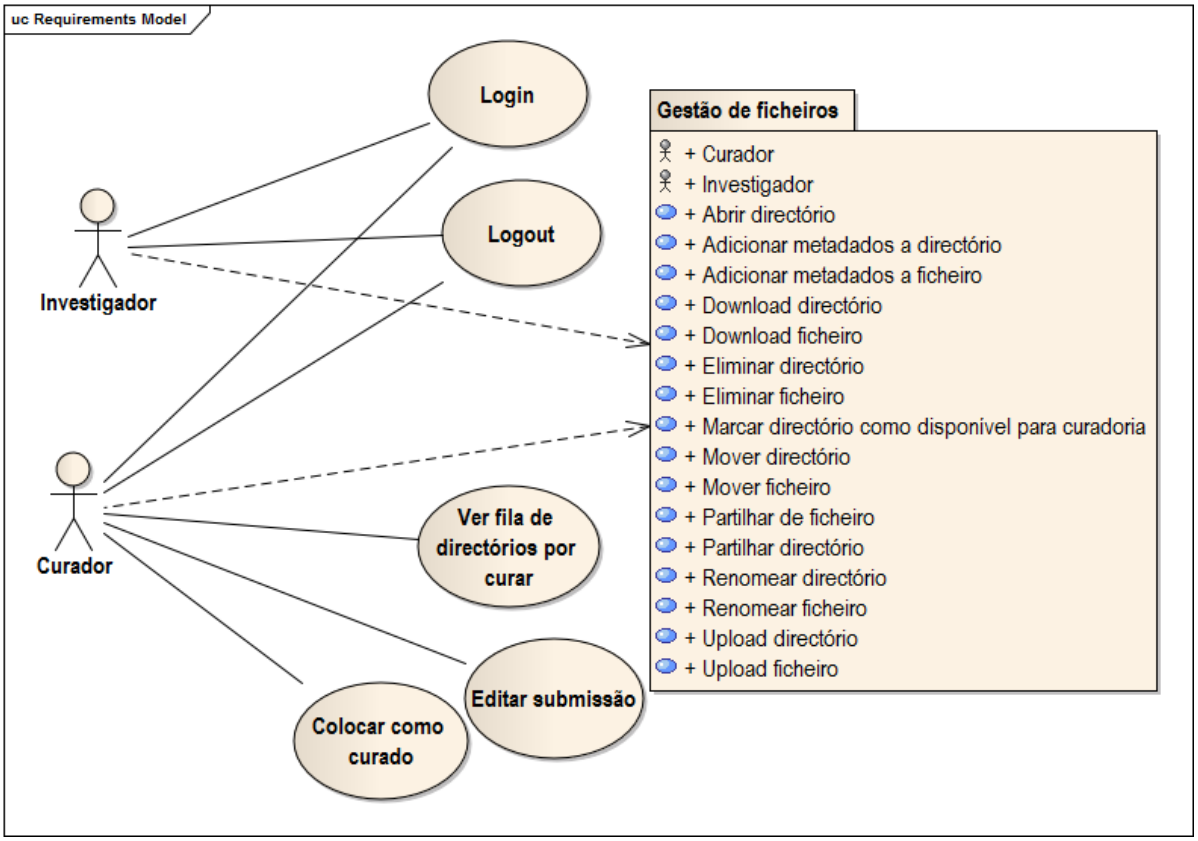

Figura 4.1: Requisitos da solução proposta.

É proposto um servidor baseado no protocolo *WebDav* que permita a edição e gestão de ficheiros remota. A escolha do protocolo recaiu sobre *WebDav* pois, ao contrário de FTP, este suporta nativamente a gestão permissões de acesso, atribuição de metadados a documentos e transferências de ficheiros.

O servidor inclui uma componente de base de dados e outra que é um sistema de ficheiros. A base de dados relacional armazena dados relativos aos utilizadores, aos ficheiros e à anotação dos mesmos. O sistema de ficheiros inclui um diretório por utilizador, cujo conteúdo só poderá ser acedido e editado pelo mesmo. Será necessário efetuar um estudo à melhor abordagem ao sistema de ficheiros, escolhendo o mais adequado a este problema.

Os módulos de gestão de ficheiros e permissão de acessos interagem com o sistema de ficheiros. Esta interação permite, entre outras ações, mover, editar, descarregar, submeter e eliminar ficheiros do servidor.

Para interação com o servidor será desenvolvida uma interface *web* e uma API que permite o acesso ao servidor por aplicações externas. Com o intuito de testar esta API, será desenvolvida uma pequena aplicação exemplo, que irá sincronizar um diretório local com a *UPBox*.

As anotações aos ficheiros serão efetuadas com auxílio de um sistema de anotação. Sempre que um investigador decidir anotar um conjunto de dados, será efetuado um pedido *RESTful* ao

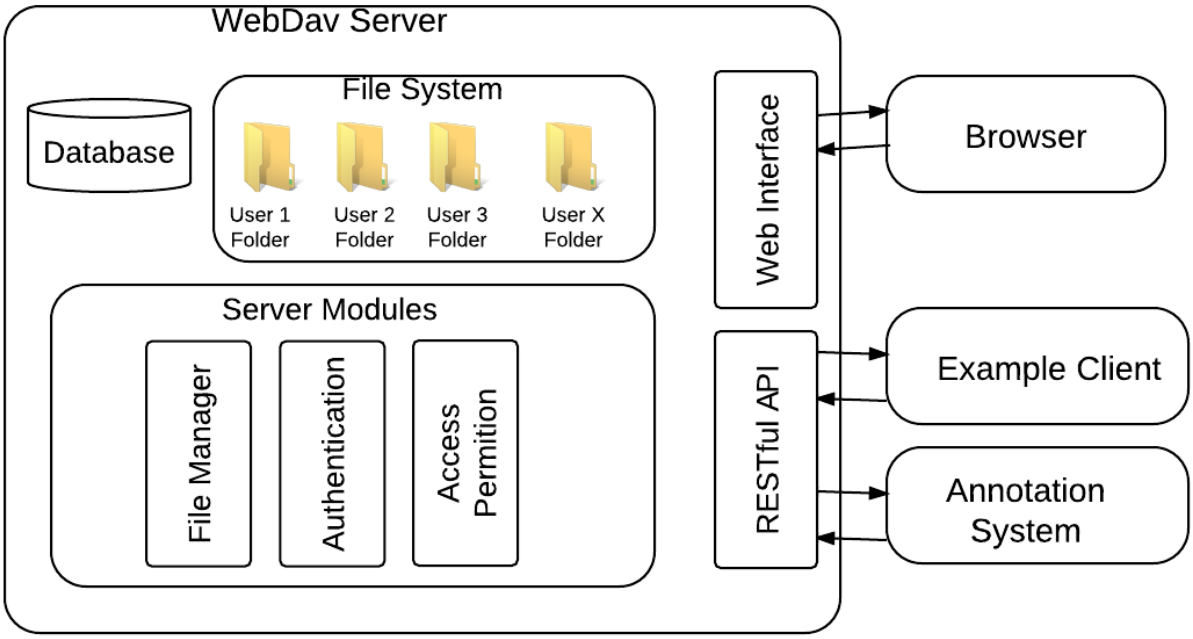

Figura 4.2: Arquitetura da solução proposta.

sistema de anotação, que irá sugerir uma anotação adequada para o *dataset*. Esta resposta será dada sob forma de um ficheiro XML com formato a definir.

Finalmente, de forma a permitir a integração de repositórios de dados com este sistema, será desenvolvida uma API que permita o acesso a ficheiros disponíveis para curadoria. Esta integração requer implementação da comunicação e receção de ficheiros nos repositórios de dados.

4.3 Metodologia e Prova de Conceito

A metodologia adotada para o desenvolvimento da UPBox será ágil iterativa, com testes unitários contínuos para validação. Uma mais valia para este projeto será o *feedback* dos investigadores que estão atualmente a colaborar com o *UPData*.

O plano de desenvolvimento da plataforma UPBox contempla o desenvolvimento de duas aplicações de prova de conceito. O objetivo é demonstrar e testar as funcionalidades da plataforma, validando a solução desenvolvida.

A primeira aplicação, e principal, será o servidor com interface *web* já referido, que será validado com testes de aceitação e através de *feedback* dos investigadores. A segunda será uma pequena aplicação exemplo que faz uso da API que permitirá a expansão da UPBox a clientes externos.

Capítulo 5

Plano de Trabalho

A elaboração da especificação da arquitetura é vital para o sucesso do projeto, por isso, como pode ser visualizado na Figura 5.1, o mês de setembro será dedicado ao *design* da aplicação. Durante esta fase será estudada a arquitetura da aplicação e definida a estrutura dos ficheiros a trocar com o sistema de anotação, e, como resultado, será produzida toda a documentação necessária à sua definição. Paralelamente serão testadas várias tecnologias de forma a escolher a mais adequada. Esta fase inclui, também, o estudo e escolha do sistema de ficheiros do sistema.

Após esta fase, de outubro a dezembro, será feita a implementação do sistema. Esta implementação será ágil e iterativa, aplicando conceitos de *design* simples. Paralelamente serão feitos testes unitários e testes de aceitação com utilizadores, de forma a testar a solução.

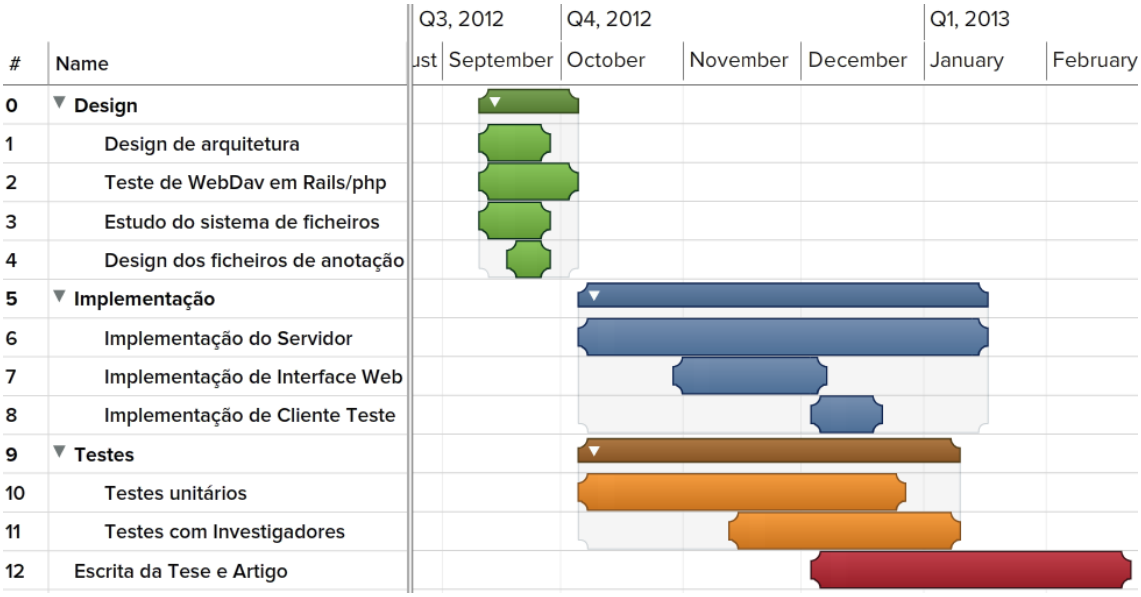

Figura 5.1: Diagrama de Gantt do projecto.

A fase de implementação inclui a implementação do servidor, das API de comunicação, da *interface web* e do cliente teste que servirá de exemplo ao funcionamento da API de expansão da UPBox a outro tipo de clientes aplicacionais.

Por fim, será escrita a dissertação e o artigo científico. O projeto esta agendado para começar em 10 de setembro de 2013 e terminar em 22 de fevereiro de 2013.

Referências

[BB04] P Bunenan e P. Buneman. The two cultures of digital curation. *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.*, pages 7–7, 2004.

[BGAT] Peter Burnhill, David Giaretta, Malcolm Atkinson e H A Tii. The Digital Curation Centre : A Vision for Digital Curation Digital Curation Centre Digital Curation Centre Edinburgh. (Dcc):31–41.

[Cen12] The Digital Curation Centre. Digital Curation Centre, 2012. [Acedido em: 22 de Junho de 2012]. Disponível em: http://www.dcc.ac.uk/digital-curation/what-digital-curation.

[CPRR11] A Data Curation, U Porto, João Rocha e Cristina Ribeiro. A Data Curation Experiment at U.Porto using DSpace. pages 224–227, 2011.

[Dat07] DataShare. DataShare Project, 2007. [Acedido em: 22 de Junho de 2012]. Disponível em: http://www.disc-uk.org/datashare.html.

[DN] F. Dridi e G. Neumann. How to implement Web-based groupware systems based on WebDAV. *Proceedings. IEEE 8th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE'99)*, pages 114–119.

[DSp12] DSpace. DSpace, 2012. [Acedido em: 22 de Junho de 2012]. Disponível em: http://www.dspace.org/introducing.

[Eco10a] The Economist. Data, data everywhere, 2010. [Acedido em: 22 de Junho de 2012]. Disponível em: http://www.economist.com/node/15557443.

[Eco10b] The Economist. The data deluge, 2010. [Acedido em: 22 de Junho de 2012]. Disponível em: http://www.economist.com/node/15579717?story_id=15579717.

[Fer06] Miguel Ferreira. Introdução à Preservação Digital: Conceitos, estratégias e actuais consensos. *Chemistry &*, 2006.

[Hey03] T Hey. The Data Deluge: An e-Science Perspective. *Grid computing*, (January 2003):1–17, 2003.

[HZZS11] Guannan Hu, Wu Zhang, Wenhao Zhu e Shijun Shen. A dynamic user-integrated cloud computing architecture. *Proceedings of the 2011 International Conference on Innovative Computing and Cloud Computing - ICCC '11*, pages 36–40, 2011.

[Lyo10] Liz Lyon. Data Dimensions : Disciplinary Differences in Research Data Sharing , Reuse and Long term Viability A comparative review based on sixteen case studies. (January), 2010.

REFERÊNCIAS

[Moh12] Arif Mohamed. A history of cloud computing, 2012. [Acedido em: 22 de Junho de 2012]. Disponível em: http://www.computerweekly.com/feature/A-history-of-cloud-computing.

[Mur12] David Murphy. Why is Dropbox Successful? It's the Simplicity, 2012. [Acedido em: 22 de Junho de 2012]. Disponível em: http://www.pcmag.com/article2/0,2817,2400746,00.asp.

[REFP11] Cristina Ribeiro, Maria Eugénia, Matos Fernandes e U Porto. Data Curation at U . Porto : Identifying current practices across disciplinary domains by. pages 14–17, 2011.

[RJKG10] Bhaskar Prasad Rimal, Admela Jukan, Dimitrios Katsaros e Yves Goeleven. Architectural Requirements for Cloud Computing Systems: An Enterprise Cloud Approach. *Journal of Grid Computing*, 9(1):3–26, December 2010.

[Rom10] Jeroen Rombouts. Building a'data repository'for heterogenous technical research communities through collaborations. 2010.

[Rou11] Margaret Rouse. Cloud Storage, 2011. [Acedido em: 22 de Junho de 2012]. Disponível em: http://searchcloudstorage.techtarget.com/definition/cloud-storage.

[RR] João Rocha e Cristina Ribeiro. Managing research data at U . Porto : requirements , technologies and services.

[RSR+10] Cristina Ribeiro, Ricardo Saraiva, Eloy Rodrigues, Matos Fernandes, Cristina Marques Gomes e José Carvalho. Os Repositórios de Dados Científicos: Estado da Arte. 2010.

[Sab11] Farzad Sabahi. Cloud computing security threats and responses. *2011 IEEE 3rd International Conference on Communication Software and Networks*, pages 245–249, May 2011.

[SCL] Fu-quan Sun, Xu Cheng e Chao Liu. Research on Hadoop-based Enterprise File Cloud.

[SL12] Martin Stephenson e Giusy Di Lorenzo. Open Innovation Portal : a collaborative platform for open city data sharing. (March):522–524, 2012.

[WPG+10] Jiyi Wu, Lingdi Ping, Xiaoping Ge, Ya Wang e Jianqing Fu. Cloud Storage as the Infrastructure of Cloud Computing. *2010 International Conference on Intelligent Computing and Cognitive Informatics*, pages 380–383, June 2010.

[WVW12] Leslie Willcocks, Will Venters e Edgar A. Whitley. Meeting the challenges of cloud computing, 2012. [Acedido em: 22 de Junho de 2012]. Disponível em: http://www.accenture.com/us-en/outlook/Pages/outlook-online-2011-challenges-cloud-computing.aspx.